Supplementary Material
Hand Pose Estimation for Pediatric Bone Age Assessment

María Escobar[1]⋆(✉), Cristina González[1]⋆, Felipe Torres[1], Laura Daza[1], Gustavo Triana[2], and Pablo Arbeláez[1]

[1] Universidad de los Andes, Bogotá, Colombia
{mc.escobar11, ci.gonzalez10, f.torres11, la.daza10, pa.arbelaez}@uniandes.edu.co
[2] Fundación Santa Fe de Bogotá, Colombia

1 Dataset Description

We provide a further description of the dataset created by the Radiological Society of North America (RSNA) for the 2017 Pediatric Bone Age Challenge [1] and a comparison between the RSNA dataset and RHPE. The RSNA dataset is composed of 14,236 images divided into 3 sets: 12,611 (88.6%) for training, 1,425 (10%) for validation and 200 (1.4%) for testing. Additionally, 46% of the RSNA dataset corresponds to female patients and 54% to male patients. We preserved the same proportion of quantity of images for each split of the RHPE dataset. We also preserved the same bone age and gender distribution throughout evey split of RHPE to ensure consistent performance in training, validation and testing. Figure 1 and 2 show the bone age distribution of the RHPE and RSNA dataset with a fitted gaussian centered around 125 and 126 months of age respectively. Figure 3 shows the direct comparison between both gaussian distributions, the red line corresponds to RSNA and the green line to RHPE. In Figure 3 we find that both distributions are similar and centered around the same age.

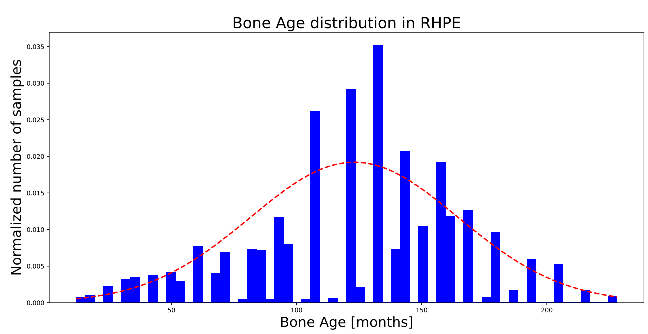

Fig. 1. Bone Age distribution in the RHPE dataset with a fitted gaussian distribution.

⋆ Both authors contributed equally to this work.

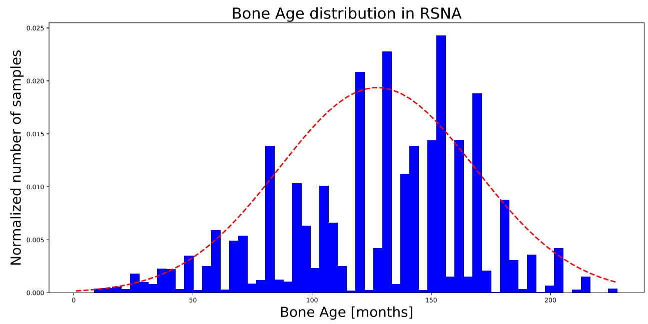

Fig. 2. Bone Age distribution in the RSNA dataset with a fitted gaussian distribution.

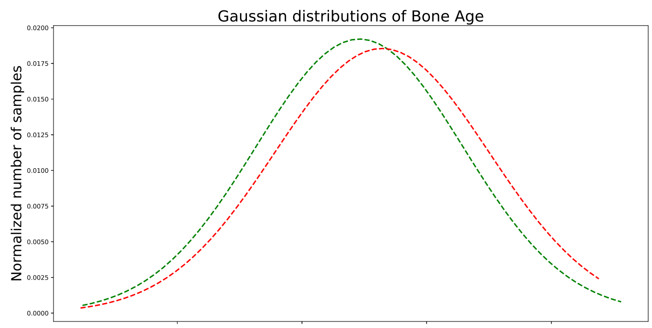

Fig. 3. Comparison between the gaussian ditribution of RHPE (green) and RSNA(red).

2 Hand Pose Estimation

The Object Keypoint Similarity (OKS) metric presented in MS-COCO [2] uses a standard deviation σ_i for every keypoint. It corresponds to the allowed deviation in localizing the keypoints. Since some regions have to be better localized than others these σ_i vary with respect to object scale. We compute the standard deviation for each keypoint by calculating the variance that different trained annotators have on the same image. Our calculated σ are one order of magnitude lower than the σ reported in MS-COCO for Human pose estimation [2]. In full-sized images, our σ penalizes any keypoint estimation 10 pixels away or more from the mean location. Table 1 shows the deviations for each of our keypoints and Figure 4 shows the location of each keypoint in the hand radiograph.

To evaluate this task we aim for a penalization of small variations in keypoint localization. Therefore, we analyze the effect that modifying the σ used in OKS has on mAP@[.5:.05:.95] for the RSNA dataset.

Table 1. Computed standard deviation σ for each keypoint taking into account the different annotations for the same image.

Keypoint	f1_0	f1_1	f1_2	f2_0	f2_1	f2_2	f3_0	f3_1	f3_2	f4_0	f4_1	f4_2	f5_0	f5_1	f5_2	anchor_0	anchor_1
σ	0.0110	0.0107	0.0090	0.0076	0.0074	0.0071	0.0039	0.0039	0.0040	0.0070	0.0071	0.0071	0.0120	0.0120	0.0120	0.0090	0.0080

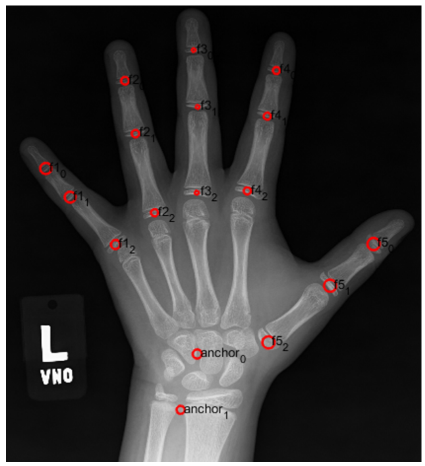

Fig. 4. Example image with circles of the computed σ denoting the deviation in the annotations for each keypoint type.

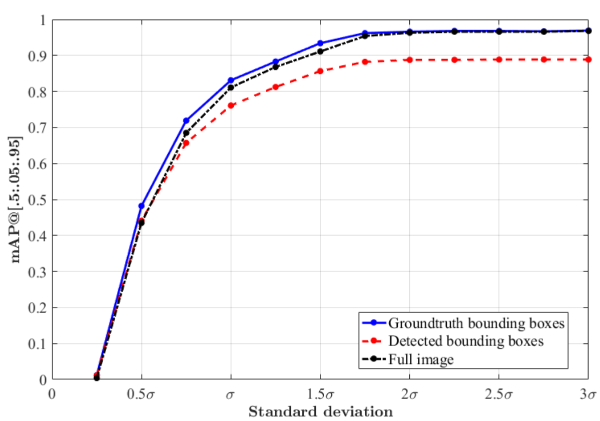

Fig. 5. Relation between σ and mAP@[.5:.05:.95] on different experiments. We use one standard deviation to achieve a strict evaluation metric.

Figure 5 shows that increasing σ for each keypoint type drives the mAP@[.5:.05:.95] to a value close to 1. However, when using σ calculated with the standard deviation of the annotations, the performance lowers. This implies that using one standard deviation for each keypoint type will be the most strict configuration for OKS.

We also provide an error diagnosis of the hand pose estimation task in the RSNA dataset following the guidelines dictated by [3]. In figure 6 we find the results of the error analysis for the localization of keypoints in hand radiographs. Here we find that solving jitter related problems would improve significantly the performance in this task.

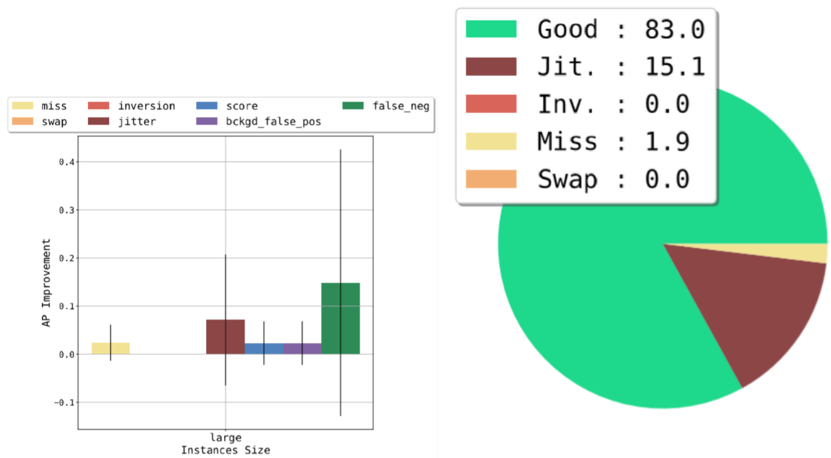

Fig. 6. Results of error analysis for the hand pose estimation task in the RSNA dataset. a) AP improvement given by fixing each type of error. b) Percentage of images with a certain error type.

3 BoNet

Figure 7 illustrates the complete architecture of BoNet. First, the hand radiograph and the anatomical ROIs heatmap are each passed through a 3x3 convolution with *stride=2*. The weights of each convolution layer are not shared during this stage. Then we concatenate both outputs and pass them through a 1x1 convolution. Subsequently, the result goes through the remaining layers of the Inception-v3 [4] architecture (CNN). The binary value of gender goes through a 32-neuron fully connected layer to generate a 32 bits embedding. Finally, the image features and the gender embedding are concatenated and go through two 1000-neuron fully connected layers to obtain the bone age prediction.

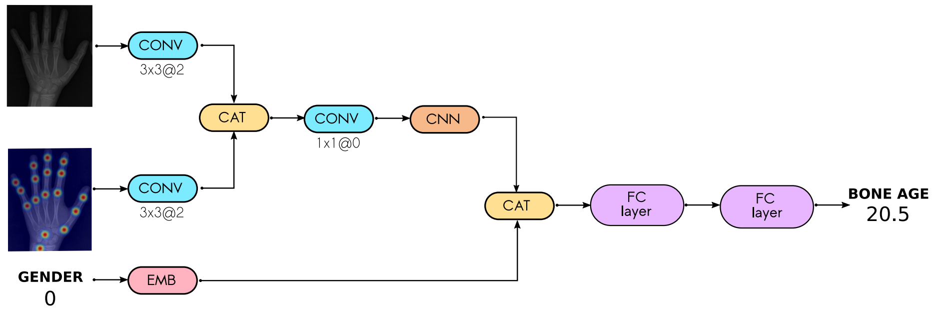

Fig. 7. Overview of the complete architecture of BoNet.

References

1. Halabi, S.S., Prevedello, L.M., et al: The rsna pediatric bone age machine learning challenge. Radiology **290**(2), 498–503 (2019). https://doi.org/10.1148/radiol.2018180736, https://doi.org/10.1148/radiol.2018180736
2. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
3. M. R. Ronchi and P. Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
4. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308,